Hessisches Statistisches Landesamt

STATISTIK HESSEN

HESSEN

URL finding: looking back, progress and plans for the future

Web Intelligence Network Conference
04.02.2025
Heidi Kühnemann
Statistics Hesse

Outline

- URL finding in a nutshell
- What has been done (during the WIN):
  - URL finding methodology report
  - WP 2 OBEC annotation exercise
- Browser comparison (at Statistics Hesse)
- Plans and ideas for the future

URL finding in a nutshell

Automated procedure to identify enterprise websites, usually containing these steps:

- sending search terms to a search engine,
- scraping the result URLs,
- extracting the relevant information from the scraped data and
- applying a machine learning or rule-based model to link websites to enterprises

URL finding at the WIN: cooperation of WP 2 OBEC & WP 3 UC 5

- Both WPs use enterprise websites as starting point
  - OBEC: classifiy online shops, social media presence, … from enterprise websites for new indicators
  - UC 5: classify economic activity codes and extract contact information from enterprise websites to enhance the business register

→ Common methodology report on URL finding

→ Systematic collection of experiences and practical advice to implement URL finding

Some challenges with URL finding

- URL finding not possible on the WIH
- Scraping search engine result URLs is costly (e.g. storage, processing capacities)
- URL finding is a difficult task, eg. because:
  - many to many relationship between enterprises and websites
  - Not all websites with enterprise data are the correct websites
  - SBR data and website data are not always identical
- Despite all this: satisfactory accuracy of URL finding

WP 2 OBEC: Manual creation of evaluation data

→ Presentation by Ville Auno

Annotation exercise for URL finding, e-commerce and social media presence

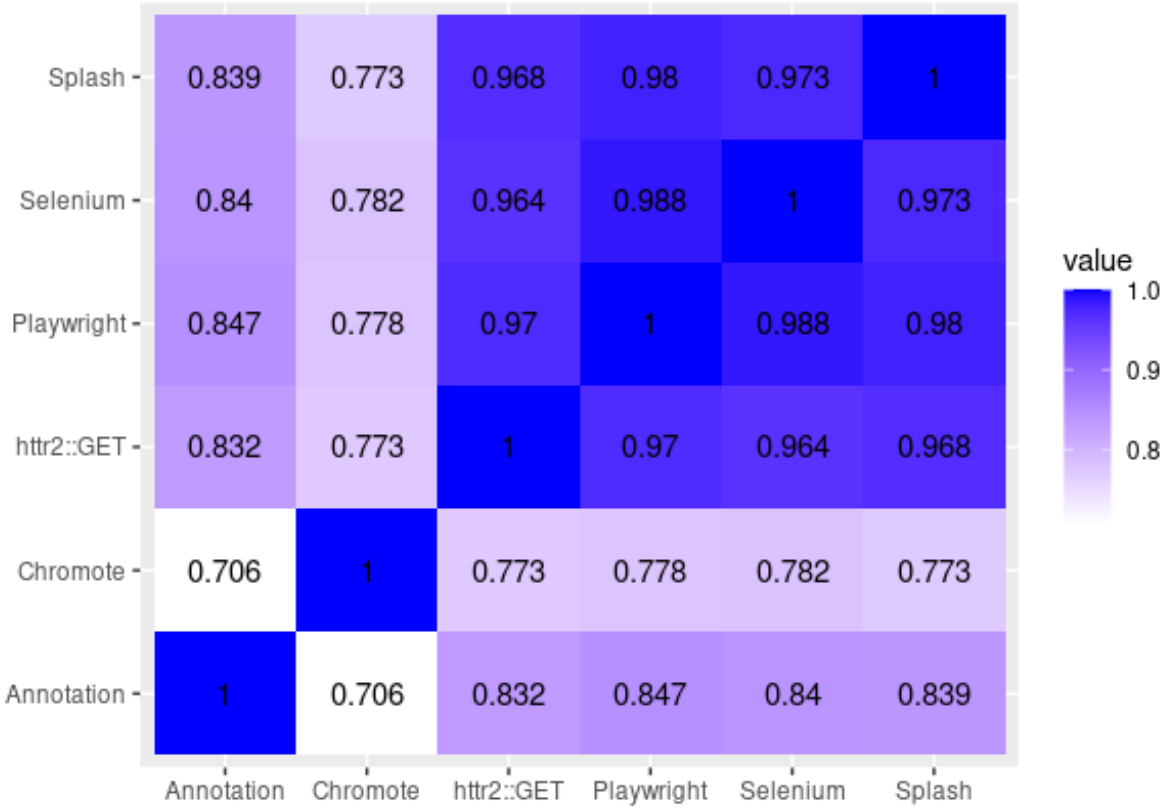
→ Common annotation handbook & evaluation script

→ Evaluation by WP 4

Country	Accuracy (%)	Weighted Accuracy (%)
Austria	87.4	86.5
Bulgaria	83.6	84.4
Germany (Hesse)	87.8	87.5
Italy	89.6	97.4
Lithuania	82.7	82.7

Percentage of correct matches between URLs found by manual annotation versus URLs found by an automated URL finding in each country (WIN Deliverable 4.8, Auno et al. 2024)

Comparing different browsers at Statistics Hesse



Comparing found URLs using different scraping methods with annotated labels (correlation of labels)

Data source: manually annotated URLs from OBEC annotation exercise

	N Companies with URL	Prop Companies with URL	N Correct URLs	Prop of correct URLs found
Annotation	**372**	**0.744**	**372**	**1**
Playwright	364	0.728	330	0.887
Selenium	358	0.716	325	0.874
Splash	357	0.714	324	0.871
httr2::GET	351	0.702	319	0.858
Chromote	265	0.53	240	0.645

Plans for the future at Statistics Hesse

- Fully automating the URL finding process
- Comparing enterprise information from imprint pages with Statistical Business Register (SBR) data using Nested Named Entity Recognition
- Procedure to regularly check and (if needed) update URLs
  - → Visit URLs, extract imprint and other data, compare to SBR using ML
  - → If URL is not correct anymore: Repeat URL finding process

Other interesting topics for the future: LLMs (→ Donato Summa) and Open Web Search Data/European Search Engine

References

Auno, Ville; Six, Magdalena; Gussenbauer, Johannes; Kowarik, Alexander (2024): **Deliverable 4.8: Quality Assessment for the Statistical Use of Web Scraped Data**. Web Intelligence Network.

Kühnemann, Heidi; van Delden, Arnout; Summa, Donato; Georgiev, Kostadin; Gussenbauer, Johannes; Ils, Alexandra; Löytynoja, Katja (2022): **Report: URL finding methodology. Joint report for Work Package 2 (Online Based Enterprise Characteristics) and Work Package 3, Use Case 5 (Business register quality enhancement).** https://cros.ec.europa.eu/system/files/2023-12/20220131_url_finding_methodology.pdf

Contact

Heidi Kühnemann

Data Scientist in section „Software Development“

Hessisches Statistisches Landesamt

Phone: +49 (0)611 3802-146

E-Mail: heidi.kuehnemann@statistik.hessen.de

https://statistik.hessen.de